



Research Report

ETS RR-11-07

Equating Subscores Using Total Scaled Scores as an Anchor

Gautam Puhan

Longjuan Liang

March 2011

Equating Subscores Using Total Scaled Scores as an Anchor

Gautam Puhan and Longjuan Liang
ETS, Princeton, New Jersey

March 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Dan Eignor

Technical Reviewers: Mary Grant and Adele (Xuan) Tan

Copyright © 2011 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING., are registered trademarks of Educational Testing Service (ETS). PRAXIS is a trademark of ETS.



Abstract

Because the demand for subscores is ever increasing, this study examined two different approaches for equating subscores: (a) equating a subscore on the new form to the same subscore in the old form using internal common items as the anchor to conduct the equating, and (b) equating a subscore on the new form to the same subscore in the old form using equated total scores as the anchor to conduct the equating. Equated total scores can be used as an anchor to equate the subscores because the total equated scores are comparable across both the new and the old forms. Data from 2 tests (Tests X and Y) were used to conduct the study, and results showed that when the number of internal common items was large (approximately 50% of the total subscore), then using common items to equate the subscores was preferable. However, when the number of common items was small (approximately 25% of the total subscore, which is common practice), then using total scaled scores (TSS) to equate the subscores was preferable. Using raw subscores (not equating) resulted in a considerable amount of bias for both tests.

Key words: subscores, test equating, anchor test, scaled scores, identity equating

Table of Contents

	Page
Introduction.....	1
Purpose of the Study	3
Method	4
Data.....	4
The Approach	4
Equating Criterion	5
Experimental Conditions for Test X.....	6
Experimental Conditions for Test Y.....	7
Procedure Followed to Evaluate the Subscore Equatings in Conditions 1 and 2.....	7
Results.....	9
Results for Test X.....	12
Results for Test Y	19
Discussion and Conclusion.....	20
Future Research	22
References.....	24
Notes	25
Appendix.....	26

List of Tables

	Page
Table 1. Summary Statistics for New, Old, and Anchor Tests (Reading, Math, Social Studies, and Science) for Test X.....	10
Table 2. Summary Statistics for New, Old, and Anchor Tests (Reading, Math, and Writing) for Test Y	11
Table 3. Weighted Average of the Conditional Standard Error of Equating (Avg SEE), Weighted Average Bias (Bias) and Root Mean Squared Deviation (RMSD) for Test X (Reading, Math, Social Studies, and Science Subscores)	13
Table 4. Weighted Average of the Conditional Standard Error of Equating (Avg SEE), Weighted Average Bias (Bias) and Root Mean Squared Deviation (RMSD) for Test Y (Reading, Math, and Writing Subscores).....	14

List of Figures

	Page
Figure 1. Graph showing two alternate subforms (X1 and X2) created from the original Form X.	5
Figure 2. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, social studies, and science subscore equatings for Test X (Condition 1).	15
Figure 3. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, social studies, and science subscore equatings for Test X (Condition 2).	16
Figure 4. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, and writing subscore equatings for Test Y (Condition 1).	17
Figure 5. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, and writing subscore equatings for Test Y (Condition 2).	18

Introduction

Testing programs typically report total test scores but not subscores to examinees and academic institutions (e.g., school districts, colleges, and universities). However, the demand for subscores is increasing due to at least two important reasons. First, examinees want to know their strengths and weaknesses in different content areas (e.g., geometry or algebra subsections on a mathematics test) to plan future remedial studies. Second, states and academic institutions such as colleges and universities want a profile of performance for their graduates to better evaluate their training and focus on areas that need remediation (Haladyna & Kramer, 2004).

Despite this increasing demand, certain important factors must be considered before subscores can be viewed as useful for diagnostic, remedial, or other high stakes purposes. For instance, subscores should be highly reliable (Sinharay, 2010). Because subscores usually are based on a much smaller number of items than the total score, however, high reliability is difficult. According to Sinharay, Haberman, and Puhan (2007), although many tests are designed to cover a broad range of abilities and the total test score is considered a composite of different abilities measured by different subsections, a subsection with fewer items than the total test may not be able to accurately measure a unique ability.

To make a meaningful interpretation of subscores, subscores must be comparable across different forms of a test. For example, the same reading subscore on an easy form would not indicate the same level of knowledge and skills as compared to that same score on a difficult form, and therefore an adjustment (e.g., equating) may be needed to make the scores comparable. Comparability of subscores across different forms of the same test is important at both the examinee and institutional (aggregate) levels. For example, *repeaters*—examinees who have failed the test and who take a different form of the test in a future test administration—should have the ability to compare subscores on both forms of the test according to a common metric. The common metric would allow repeaters to assess their progress on some or all subscores. Similarly, institutions such as colleges and universities often need annual summaries of subscores for their candidates to plan remedial training programs. However, aggregating subscores from different forms of the same test is only meaningful if the subscores are comparable across different forms of the test. Finally, comparability of subscores across different forms of a test is extremely important if subscores are used either fully or in conjunction with the total equated scores to make high stakes decisions such as hiring or certification.

Both high reliability and comparability of subscores across different forms are important factors to consider before reporting subscores may be justified, this study focuses only on different approaches for equating subscores to make them comparable across different forms of the same test.¹

Equating methods are commonly used to adjust difficulty differences across parallel forms of a test, which in turn allow for score comparisons across different groups of examinees regardless of the test forms they were administered. Although guidelines for equating total scores under different equating designs such as the single group equating design, common item equating design, etc., can be found in the equating literature (see Livingston, 2004; Kolen & Brennan, 2004), similar guidelines for equating subscores are not well documented. Because subscores have fewer items than the total score, equating subscores using designs such as the common item equating design may be difficult if the number of common items needed to conduct the subscore equating are too small. Consider a test consisting of 100 items where approximately 25 items (typical practice) are used as common items to equate the new and old forms of the test. If this test has five subscores with 20 items in each subscore, then approximately five items in each subscore can be used as common items to equate a particular subscore (e.g., mathematics) in the new form to the same subscore in the old form. With such a small common item set, it may be difficult to adequately represent the subscore in terms of content and difficulty, which in turn may lead to a less precise equating (Livingston, 2004).

Although some attempts have been made under the item response theory (IRT) framework to report subscores on a common metric across different forms of a test (e.g., Yao & Boughton, 2006), very little research exists on equating subscores in an observed score equating framework (e.g., using equipercentile or linear equating methods). Furthermore, although some testing programs report subscores that are equated using observed score equating methods (e.g., ETS[®] Proficiency Profile), no studies exist that show that the current approaches to equate these subscores are indeed adequate. Considering the demand for subscores is constantly rising, an empirical evaluation of procedures for effectively equating subscores becomes especially important.

Purpose of the Study

The purpose of this study is to compare different approaches to make subscores comparable across different forms of the same test. The two approaches to be examined are described as follows.

1. A subscore (e.g., reading) on the new form will be equated to the same subscore on the old form using internal common items as the anchor to conduct the equating.
2. A subscore on the new form will be equated to the same subscore on the old form using equated scaled total scores as the anchor to conduct the equating. Because total scaled scores (TSS) are comparable across the new and old forms, they can be used as an anchor to equate the subscores.²

The two approaches will be compared to a criterion subscore equating to evaluate which approach leads to a more precise subscore equating. For informational purposes, raw subscores will also be compared to the criterion to evaluate if reporting raw subscores is reasonable. When raw total or subscores are reported, a strong assumption is made that the new and old forms of the test are comparable.

The working hypothesis is that using equated total scores as an anchor may lead to a more precise equating as compared to using internal common items as an anchor, especially in conditions where the total score is moderately to highly correlated to the subscores and/or when very few items exist that can be used as an internal anchor to equate subscores. We acknowledge that when the total score is used as an anchor to equate subscores, it is implicitly assumed that the subscores and the total score measure the same knowledge and skills; therefore, reporting subscores in such situations may seem redundant. However, as pointed out in Puhan, Sinharay, Haberman, and Larkin (2010), in situations where subscores may appear redundant (i.e., as evident by a high correlation between the subscore and the already reported total score) yet not harmful (i.e., as evident by a high subscore reliability), it may be reasonable to report subscores as they have a perceived usefulness for users. In situations where the subscores truly represent distinct dimensions (e.g., reading, mathematics), equating through the total scores may not be a good choice. A possible yet somewhat impractical option may be to have the new and old forms overlap considerably (e.g., 50% to 60% overlap) so enough common items exist between two subscores across the new and old forms to equate them effectively under the common item equating model. A second option (if sample sizes for the test under consideration are large) is to

spiral in alternating sequence the new form of the test with an old form of the test, which is already on scale. Then the subscores on the new form can be equated to the subscores on the old form using the randomly equivalent groups design without using common items.

Method

Data

Data from two operational tests, referred to as Tests X and Y, were used in this study. Data for Test X consisted of 23,418 examinee responses from one form of Test X (i.e., Form X). These 23,418 examinee responses were accumulated from four different test administrations in which Form X was administered without any changes (i.e., a reprint form). Data for Test Y consisted of 4,815 examinee responses from one form of Test Y (i.e., Form Y). These 4,815 examinee responses were accumulated from two different test administrations in which Form Y was administered without any changes. Test X consisted of 120 multiple-choice items measuring basic skills in elementary education and covered four subcontent areas, or subscores (i.e., reading, mathematics, social studies, and science). Each subcontent area contributed equally to the total test (i.e., 30 out of 120 items, or 25% of the total test). Test Y consisted of 90 multiple-choice items measuring basic skills required of teacher aides and covered three subcontent areas (i.e., reading, mathematics, and writing). Each subcontent area contributed equally to the total test (i.e., 30 out of 90 items or 33.33% of the total test).

The Approach

For the purpose of this study, a hypothetical testing situation was created whereby the single Form X was divided into two pseudoforms (i.e., Forms X1 and X2) with 84 items in each form and with 48 items common to both forms (see Figure 1 for illustration). As seen in Figure 1, Form X (consisting of 120 items) is divided into two alternate subforms, Forms X1 and X2, each consisting of 84 items. The shaded portion indicates the common section of 48 items between Forms X1 and X2. This procedure of using one original form to build two subforms and conducting the equating on the subforms allows for the creation of a strong equating criterion, which is discussed in the next section. The items for Forms X1 and X2 were selected in a way such that Forms X1 and X2 comprised the four content areas of reading, mathematics, social studies, and science, with each content area contributing to 25% of the total test. However, because Forms X1 and X2 now had fewer items than Form X, the number of items in each

content area was fewer than Form X (i.e., 21 instead of 30 items in each subcontent area). Throughout the study, Form X1 will be considered the new form and Form X2 will be considered the old form.

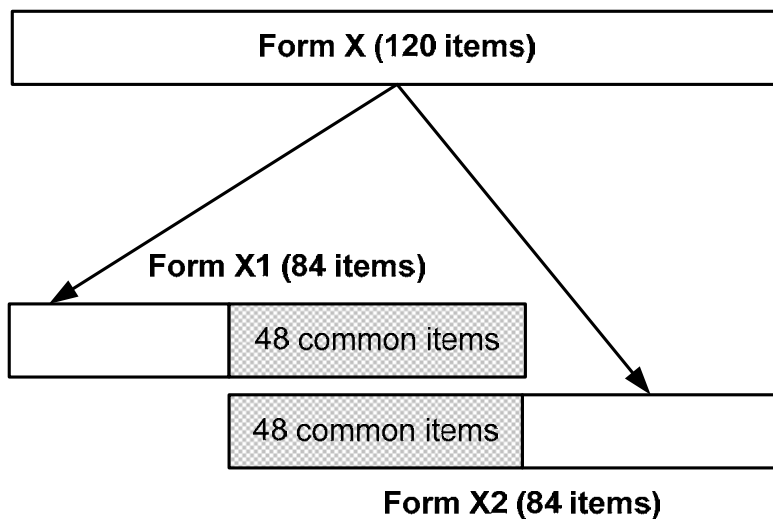


Figure 1. Graph showing two alternate subforms (X1 and X2) created from the original Form X.

The same approach was used to create two pseudoforms for Form Y, where Form Y (consisting of 90 items) was divided into two alternate subforms, Forms Y1 and Y2, each consisting of 60 items with 30 common items between them. The items for Forms Y1 and Y2 were selected in a way such that Forms Y1 and Y2 comprised the three content areas of reading, mathematics, and writing, with each content area contributing to 33.33% of the total test. However, because Forms Y1 and Y2 now had fewer items than Form Y, the number of items in each content area was fewer than Form Y (i.e., 20 instead of 30 items in each subcontent area).

Equating Criterion

The hypothetical test design described above facilitated the computing of a criterion equating function. Because Forms X1 and X2 were created from one Form X (taken by 23,418 examinees), essentially all the examinees took Form X1 and also Form X2. Therefore, a particular subscore (e.g., reading) on Form X1 could be equated directly to the same subscore on Form X2 using a single group (SG) equating design. By using the SG design with such a large data set, one can be fairly confident that the resulting conversion is a very good approximation of

the equating in the population. Livingston (1993) first proposed this method for deriving a criterion equating function for the total test. The same approach was followed for Test Y. Since Forms Y1 and Y2 were created from Form Y (taken by 4,815 examinees), essentially all examinees took Form Y1 and Y2, and therefore, a particular subscore (e.g., mathematics) on Form Y1 can be equated to the same subscore on Form Y2 using an SG equating design.

Experimental Conditions for Test X

For Test X the subscore equatings were examined under two conditions. Condition 1 was identical to the hypothetical situation shown in Figure 1 where a large overlap existed between Forms X1 and X2 (i.e., 48 common items could be used as equaters). In Condition 2, a smaller overlap between Forms X1 and X2 was assumed. Although the actual overlap between Forms X1 and X2 was 48 items, only 24 of these common items were identified as equaters to mimic actual testing situations where the common items constitute approximately 20% to 25% of the total test. Both common items sets (i.e., 48 or 24 items) were created to be miniversions of the total test.

It was hypothesized that if the subscores and the total score were moderately to highly correlated, then in Condition 1, one wouldn't expect to see much difference in the subscore equatings derived using either internal common items or TSS as the anchor. Because 48 common items were present at the total test level, enough common items were present at the subscore level (i.e., 12 out of 21 items for each subscore) to result in a fairly accurate equating. Similarly, if subscores were moderately to highly correlated to the total scores, then using TSS as an anchor may also result in an accurate equating. However, if the test is truly multidimensional, then it is expected that using common items to conduct the subscore equatings will result in a more accurate equating as compared to using TSS to conduct the subscore equatings.

In Condition 2, where fewer common items were assumed at the total test score level, the number of common items that could be used to equate the subscores was much smaller. For example, when Forms X1 and X2 have 24 common items, then the approximate number of items that could be used as common items to equate a particular subscore (e.g., reading) in Form X1 to the reading subscore in Form X2 was 6 items. Such a small number of internal common items may result in an imprecise equating as compared to equating the subscores using TSS. Under such circumstances, equating subscores using TSS may prove to be more beneficial.

Experimental Conditions for Test Y

For Test Y, the subscore equatings were also examined under two conditions. In Condition 1, a large overlap existed between Forms Y1 and Y2 (i.e., 30 common items). In Condition 2, a smaller overlap between Forms Y1 and Y2 was assumed (i.e., 15 common items). Both common items sets (i.e., 30 or 15 items) were created to be miniversions of the total test.

Similar to Test X, it was hypothesized that if the subscores and the total score are moderately to highly correlated, then in Condition 1, one wouldn't expect to see much difference in the subscore equatings derived using either internal common items or TSS as the anchor. Because 30 common items were present at the total test level, enough common items were present at the subscore level (i.e., 10 out of 20 items for each subscore) to result in a fairly accurate equating. Similarly, if subscores were moderately to highly correlated to the total scores, then using the TSS may also result in a similar equating.

In Condition 2 where fewer common items are assumed at the total test score level, the number of common items that could be used to equate the subscores was much smaller. For example, when Forms Y1 and Y2 have 15 common items, then the approximate number of items that could be used as common items to equate a particular subscore (e.g., mathematics) in Form Y1 to the same subscore in Form Y2 was 5 items. Such a small number of internal common items may result in an imprecise equating as compared to equating the subscores using TSS.

Procedure Followed to Evaluate the Subscore Equatings in Conditions 1 and 2

The two approaches (i.e., using internal common items as an anchor to equate subscores on the new and old forms and using TSS as an anchor to equate subscores on the new and old forms) were compared to the criterion equating. For informational purposes, raw subscores were also compared to the criterion to evaluate if reporting raw subscores was a reasonable approach. Note that under the second approach (i.e., TSS approach), the total scores on Form X1 were first equated to the total scores on Form X2 using the common item equating design to obtain equated total scores that were subsequently used as an anchor to equate the subscores. Similarly for Test Y, the total scores on Form Y1 were first equated to the total scores on Form Y2 to obtain equated/scaled total scores that were subsequently used as an anchor to equate the subscores. For Test X (Condition 1), the total scores were equated using 48 common items as the anchor, and in Condition 2, the total scores were equated using 24 common items as the anchor. For Test Y (Condition 1) the total scores were equated using 30 common items as the anchor, and in

Condition 2, the total scores were equated using 15 common items as the anchor. The study was conducted using the following steps for Test X (Conditions 1 and 2). The same steps were followed for Test Y (Conditions 1 and 2).

Step 1. To estimate a criterion subscore equating function (e.g., reading), the reading subscore on Form X1 was directly equated to the reading subscore on Form X2 using the SG equating design with the total data ($N = 23,418$). The resulting conversion was considered as the criterion to which the two types of subscore equatings and no equating was compared. The criterion equatings were derived using both linear and nonlinear methods; therefore, the linear subscore equatings were compared to the linear criterion and the nonlinear subscore equatings were compared to the nonlinear criterion to avoid any confounding effect that might have arisen by comparing across linear and nonlinear equating methods. The same approach was followed for the remaining subscores (i.e., mathematics, social studies, and science). The subscore equatings that used either common items or TSS as the anchor followed the chained equating method (for details on chained equating, see Kolen & Brennan, 2004).

Step 2. For a particular condition (e.g., Condition 1), sample sizes of 1,000 each were drawn with replacement from the Form X1 and Form X2 data sets. As mentioned earlier, the total sample of 23,418 examinee responses was accumulated over four different test administrations in which the original Form X was administered without any changes. Thus, to mimic real testing conditions, examinee responses from one test administration ($N = 6,580$) were assigned to Form X1 and examinee responses from another test administration ($N = 6,798$) were assigned to Form X2; the resampling involved drawing examinee responses from these two test administrations. This measure was also true for Test Y where examinee responses from one test administration ($N = 2,424$) were assigned to Form Y1 and examinee responses from another test administration ($N = 2,391$) were assigned to Form Y2; the re-sampling involved drawing examinee responses from these two test administrations. Then the two subscore equatings were conducted (i.e., using internal common items as an anchor to equate subscores on the new and old forms and using TSS as an anchor to equate subscores on the new and old forms).

Step 3. Step 2 was repeated 500 times, and based on the repeated samples, the conditional standard error of equating (CSEE), weighted average of the CSEE (Avg SEE), conditional bias (CBias), and weighted average bias (Bias) were computed. When computing the Avg SEE and Bias, the CSEEs and CBias values were appropriately weighted using the raw

proportion of examinees at each score point in the new form data. The root mean squared deviation (RMSD) was also calculated at the total test score level for evaluating the two subscore equating approaches. The RMSD is a useful statistic because it provides an estimate based on combining information from random and systematic error. Details on these statistical indexes are provided in the appendix.

Although the different equating methods were compared relative to each other, the practical criterion of the *difference that matters* (DTM; Dorans & Feigenbaum, 1994) was also used to evaluate the different subscore equating approaches. The DTM is a unit equal to one-half of a reporting unit. Because the scores progressed in 1-point increments for the tests used in this study, the DTM was defined as 0.5. Using a DTM criterion seemed reasonable, because if a difference existed between the variability and accuracy indexes obtained using the two subscore approaches, but the actual values were smaller than the DTM, then the differences would probably be ignorable as they might not result in a practical difference in the examinees' reported scores.

Results

The summary statistics for Test X (new and old forms) and Test Y (new and old forms) for Conditions 1 and 2 are presented in Tables 1 and 2, respectively. For Test X, 48 items were used as common items in Condition 1. Therefore 12 of those 48 items were used as common items to equate each subscore (i.e., reading, math, social studies, and science). In Condition 2, 24 items were used as common items. Therefore, only 6 of those 24 items were used as common items to equate each subscore. For Test Y, 30 items were used as common items in Condition 1. Therefore 10 of those 30 items were used as common items to equate each subscore (i.e., reading, math, and writing). In Condition 2, 15 items were used as common items. Therefore, only 5 of those 15 items could be used as common items to equate each subscore.

As seen in Table 1, when internal common items were used as an anchor, the anchor-to-total test correlations for the new and old forms for Test X were fairly high in Condition 1 (min = 0.853; max = 0.935). When TSS were used as an anchor, the anchor-to-total test correlations for the new and old forms were moderately high (min = 0.733; max = 0.842). In Condition 2, when internal common items were used, the anchor-to-total test correlations for the new and old forms for Test X were moderately high (min = 0.680; max = 0.841). When TSS were used as an

Table 1

Summary Statistics for New, Old, and Anchor Tests (Reading, Math, Social Studies, and Science) for Test X

10	Condition 1	Score distributions (# of items)	R total (21)	R CI anchor (12)	R TSS anchor (84)	M total (21)	M CI anchor (12)	M TSS anchor (84)	SS total (21)	SS CI anchor (12)	SS SS anchor (84)	S total (21)	S CI anchor (12)	S TSS anchor (84)
		New form (<i>N</i> = 6,580)												
		Mean	18.13	10.26	54.81	14.03	7.95	54.81	12.65	6.69	54.81	11.07	5.57	54.81
		SD	2.27	1.53	10.38	4.20	2.61	10.38	3.09	2.01	10.38	3.42	2.27	10.38
		Anchor/total correlation		0.897	0.733		0.935	0.839		0.870	0.764		0.896	0.815
		Old form (<i>N</i> = 6,798)												
		Mean	17.31	10.13	53.98	13.73	7.82	53.98	11.21	6.58	53.98	11.73	5.43	53.98
		SD	2.56	1.56	10.21	4.05	2.59	10.21	3.05	1.99	10.21	3.31	2.26	10.21
		Anchor/total correlation		0.862	0.743		0.933	0.842		0.853	0.737		0.879	0.801
	Condition 2	Score distributions (# of Items)	R total (21)	R CI anchor (6)	R TSS anchor (84)	M total (21)	M CI anchor (6)	M TSS Anchor (84)	SS total (21)	SS CI anchor (6)	SS TSS anchor (84)	S total (21)	S CI anchor (6)	S TSS anchor (84)
		New form (<i>N</i> = 6,580)												
		Mean	18.13	5.37	54.62	14.03	4.01	54.62	12.65	3.45	54.62	11.07	2.28	54.62
		SD	2.27	0.86	10.28	4.20	1.52	10.28	3.09	1.25	10.28	3.42	1.28	10.28
		Anchor/total correlation		0.723	0.733		0.838	0.839		0.685	0.764		0.697	0.815
		Old form (<i>N</i> = 6,798)												
		Mean	17.31	5.31	53.98	13.73	3.93	53.98	11.21	3.39	53.98	11.73	2.27	53.98
		SD	2.56	0.92	10.21	4.05	1.52	10.21	3.05	1.22	10.21	3.31	1.27	10.21
Anchor/total correlation			0.701	0.743		0.841	0.842		0.680	0.737		0.689	0.801	

Note. R = reading; M = math; SS = social studies; S = science; CI = common items; TSS = total scaled scores.

Table 2

Summary Statistics for New, Old, and Anchor Tests (Reading, Math, and Writing) for Test Y

11	Condition 1	Score distributions (# of items)	R total (20)	R CI anchor (10)	R TSS anchor (60)	M total (20)	M CI anchor (10)	M TSS anchor (60)	W total (20)	W CI anchor (10)	W TSS anchor (60)
		New form (<i>N</i> = 2,424)									
		Mean	14.01	7.41	38.76	12.97	6.01	38.76	13.44	6.20	38.76
		SD	4.26	2.12	11.09	4.10	2.35	11.09	4.12	2.16	11.09
		Anchor/total correlation		0.912	0.905		0.927	0.891		0.906	0.908
		Old form (<i>N</i> = 2,391)									
		Mean	14.37	7.48	39.18	11.97	6.13	39.18	12.83	6.23	39.18
		SD	3.85	2.09	11.05	4.44	2.36	11.05	4.05	2.17	11.05
	Anchor/total correlation		0.910	0.891		0.924	0.900		0.911	0.893	
	Condition 2	Score distributions (# of items)	R total (20)	R CI anchor (5)	R TSS anchor (60)	M total (20)	M CI anchor (5)	M TSS anchor (60)	W total (20)	W CI anchor (5)	W TSS anchor (60)
		New form (<i>N</i> = 2,424)									
		Mean	14.01	3.81	38.82	12.97	3.14	38.82	13.44	3.51	38.82
		SD	4.26	1.13	11.15	4.10	1.31	11.15	4.12	1.20	11.15
		Anchor/total correlation		0.771	0.905		0.801	0.891		0.754	0.908
		Old form (<i>N</i> = 2,391)									
		Mean	14.37	3.85	39.18	11.97	3.18	39.18	12.83	3.52	39.18
SD		3.85	1.13	11.05	4.44	1.31	11.05	4.05	1.20	11.05	
Anchor/total correlation		0.770	0.891		0.798	0.900		0.769	0.893		

Note. R = reading; M = math; W = writing; CI = common items; TSS = total scaled scores.

anchor, the anchor-to-total test correlations for the new and old forms were also moderately high (min = 0.733; max = 0.842), although in many cases, the anchor-to-total correlations were higher for the TSS anchors than the common item anchors.

As seen in Table 2, when internal common items were used as an anchor, the anchor-to-total test correlations for the new and old forms for Test Y were fairly high in Condition 1 (min = 0.906; max = 0.927). When TSS were used as an anchor, the anchor-to-total test correlation for the new and old forms were also fairly high (min = 0.891; max = 0.908). In Condition 2, when internal common items were used, the anchor-to-total test correlations for the new and old forms for Test Y were moderately high (min = 0.754; max = 0.801). When TSS were used as an anchor, the anchor-to-total test correlations for the new and old forms were fairly high (min = 0.891; max = 0.908). In this condition, using TSS as compared to using internal common items as an anchor clearly resulted in a higher anchor-to-total correlation, which may result in a more accurate equating of the subscores. Note that in Tables 1 and 2, the mean and SD for the TSS anchor for the new form are not the same in Conditions 1 and 2 because they were based on different conversion lines (i.e., one was based on a longer anchor and the other was based on a shorter anchor). However, the mean and SD for the old form in both conditions are the same because they were based on the raw scores and therefore remained unchanged in both conditions.

The overall accuracy and variability of the subscore equatings using internal common items or TSS as an anchor were estimated using the Avg SEE, Bias, and RMSD indexes. These results for Tests X and Y are presented in Tables 3 and 4, respectively. The CSEEs and CBias estimates for the subscore equatings using internal common items or TSS as an anchor were also estimated. This information for Tests X and Y is presented in Figures 2 to 5. Although both chained linear and chained equipercentile equatings were conducted in this study, the results of the chained equipercentile equating did not provide much additional information than what was already observed in the chained linear equating results. Therefore, for economy of presentation, only the chained linear results are presented.³ Test X results are presented first, followed by results from Test Y.

Results for Test X

Average standard errors of equating (Avg SEE), weighted average bias (Bias), and root mean squared deviation (RMSD) results. As seen in Table 3 in Condition 1 (i.e., where a larger number of common items was used), the Avg SEE for the subscore equatings (i.e., the

Table 3

Weighted Average of the Conditional Standard Error of Equating (Avg SEE), Weighted Average Bias (Bias) and Root Mean Squared Deviation (RMSD) for Test X (Reading, Math, Social Studies, and Science Subscores)

		Reading		Math		Social studies		Science	
Condition 1	Score distributions (# of items)	CI anchor (12)	TSS anchor (84)	CI anchor (12)	TSS anchor (84)	CI anchor (12)	TSS anchor (84)	CI anchor (12)	TSS anchor (84)
	Avg SEE	0.087	0.127	0.091	0.147	0.099	0.130	0.099	0.129
	Bias	0.032	0.055	0.020	0.161	0.016	0.119	0.062	0.135
	RMSD	0.128	0.208	0.131	0.265	0.142	0.219	0.153	0.234
	Score distributions (# of items)	CI anchor (6)	TSS anchor (84)	CI anchor (6)	TSS anchor (84)	CI anchor (6)	TSS anchor (84)	CI anchor (6)	TSS anchor (84)
Condition 2	Avg SEE	0.124	0.126	0.142	0.145	0.142	0.129	0.154	0.128
	Bias	0.014	-0.008	0.030	0.061	-0.003	0.045	-0.128	0.054
	RMSD	0.208	0.186	0.208	0.215	0.202	0.192	0.253	0.189

Note. Smaller numbers are highlighted in gray. CI = common items; TSS = total scaled scores.

reading, math, social studies, and science subscores) that used internal common items as an anchor was always smaller than the subscore equatings that used TSS as an anchor. This finding was also true for the Bias and RMSD indexes. This result was not completely unexpected, because as seen in Table 1, the anchor-to-total correlations for each subscore equating were higher when common items as compared to TSS were used as the equating anchor. However, in Condition 2, the pattern was not as systematic. The Avg SEEs for the reading and math subscore equatings that used common items as an anchor were very similar to Avg SEEs for the reading and math subscore equatings that used TSS as an anchor. However, for the social studies and science subscore equatings, using common items as compared to TSS as the anchor resulted in higher Avg SEEs. The Bias values for the math and social studies subscore equatings were smaller for the common item equatings as compared to the equatings that used TSS as an anchor. However, the Bias values for the reading and science equatings were lower when TSS as

Table 4

Weighted Average of the Conditional Standard Error of Equating (Avg SEE), Weighted Average Bias (Bias) and Root Mean Squared Deviation (RMSD) for Test Y (Reading, Math, and Writing Subscores)

		Reading		Math		Writing	
		CI	TSS	CI	TSS	CI	TSS
		anchor (10)	anchor (60)	anchor (10)	anchor (60)	anchor (10)	anchor (60)
Condition 1	Avg SEE	0.104	0.109	0.103	0.119	0.101	0.113
	Bias	-0.008	-0.026	0.022	0.078	0.053	-0.048
	RMSD	0.148	0.161	0.153	0.186	0.154	0.169
Condition 2		CI	TSS	CI	TSS	CI	TSS
		anchor (5)	anchor (60)	anchor (5)	anchor (60)	anchor (5)	anchor (60)
		0.157	0.103	0.160	0.120	0.162	0.112
		-0.022	-0.006	0.106	0.103	0.069	-0.028
	RMSD	0.230	0.146	0.250	0.198	0.239	0.165

Note. Smaller numbers are highlighted in gray. CI = common items; TSS = total scaled scores.

compared to internal common items were used as the anchor. Finally, the RMSD was lower for the reading, social studies, and science equatings when TSS was used as the anchor, while the RMSD for the math equating was lower when internal common items were used as the anchor. Overall, the equatings using TSS as an anchor were more accurate (i.e., lower RMSD) for most of the subscores. These results were also not completely unexpected as the anchor-to-total correlations were either very similar or higher when TSS as compared to internal common items were used in the equating.

Conditional standard errors of equating (CSEE) and weighted average bias (Bias).

Although the average statistics described above provide a useful summary of random and systematic equating error, the CSEEs and CBias values are often considered more informative because they indicate the amount of variability and accuracy at each score point. As seen in Figure 2 for Test X (Condition 1), the CSEEs for the internal common item equating were

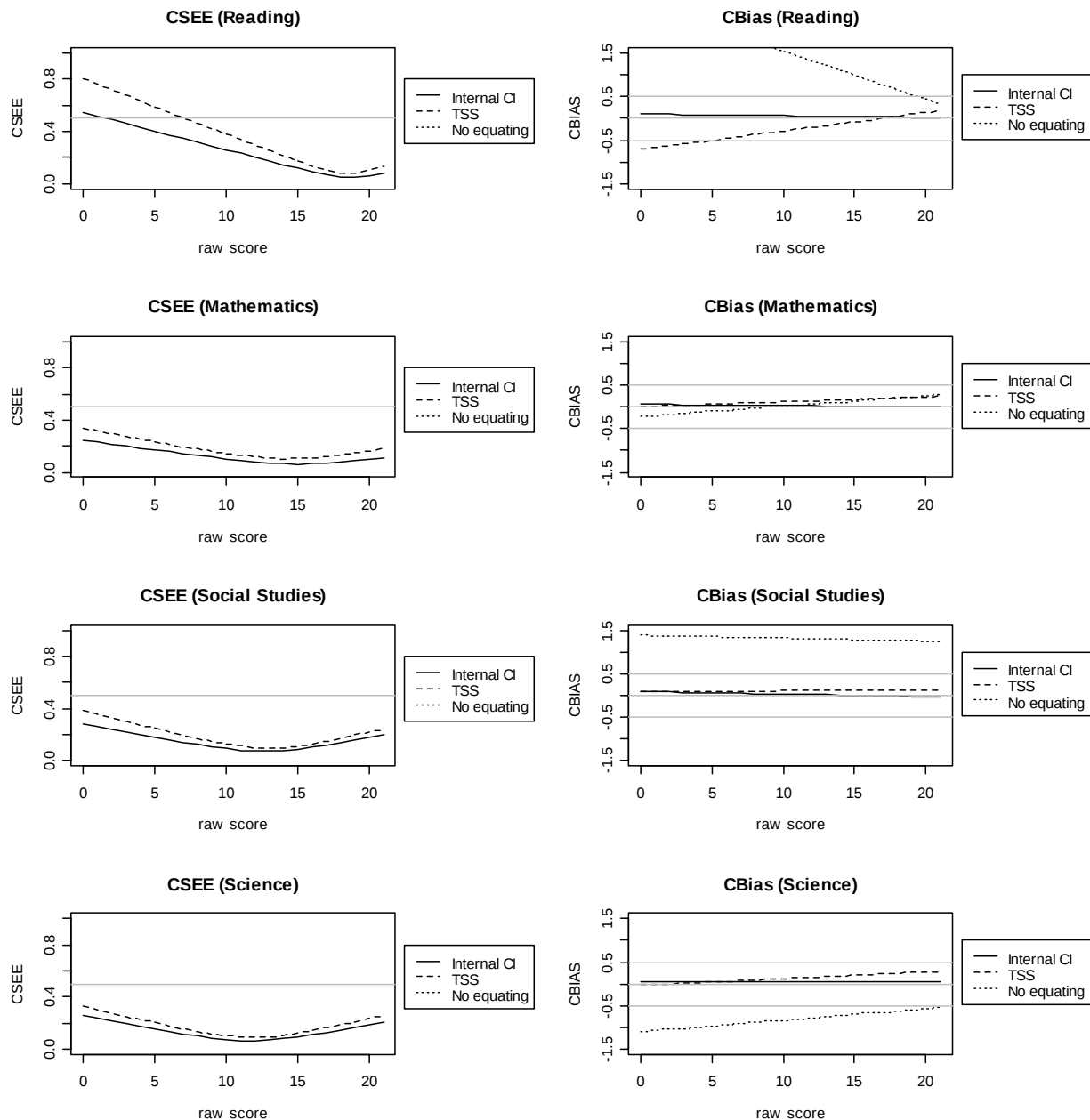


Figure 2. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, social studies, and science subscore equatings for Test X (Condition 1).

Note. The CSEE for no equating is zero and therefore not observed in the graph.

smaller than the CSEEs for the TSS anchor equating for all the subscores. However, it should be noted that the CSEEs for these subscore equatings were smaller than the DTM of 0.5 for all subscores (except for the lower score regions of the reading subscore) and therefore could be

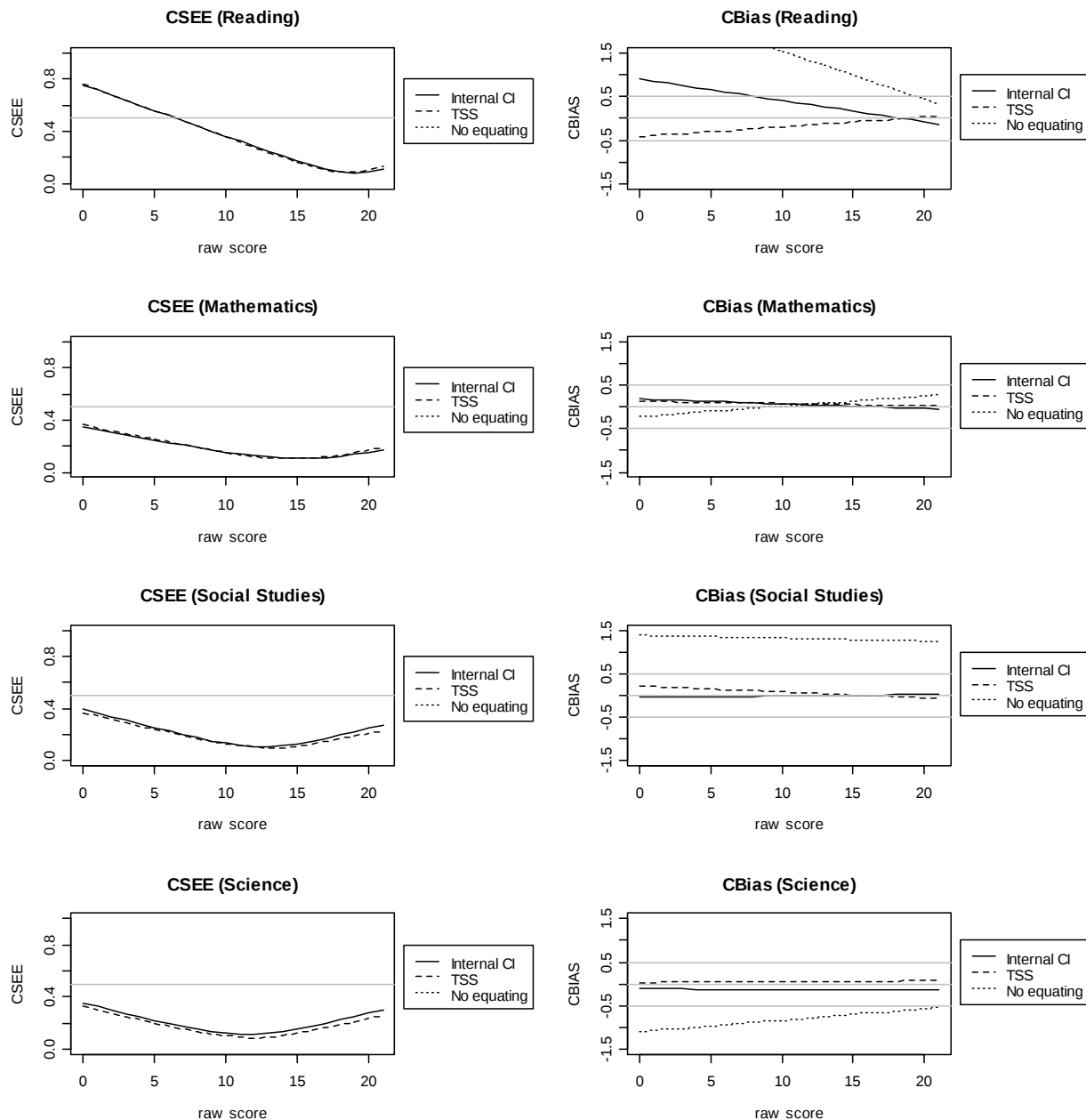


Figure 3. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, social studies, and science subscore equatings for Test X (Condition 2).

Note. The CSEE for no equating is zero and therefore not observed in the graph.

considered small. As seen in Figure 2, the CBias values for the internal common item equating and TSS anchor equating were quite small and within the DTM band for most of the score scale for all the subscores. The Bias for no equating (i.e., using raw and unequated subscores) was

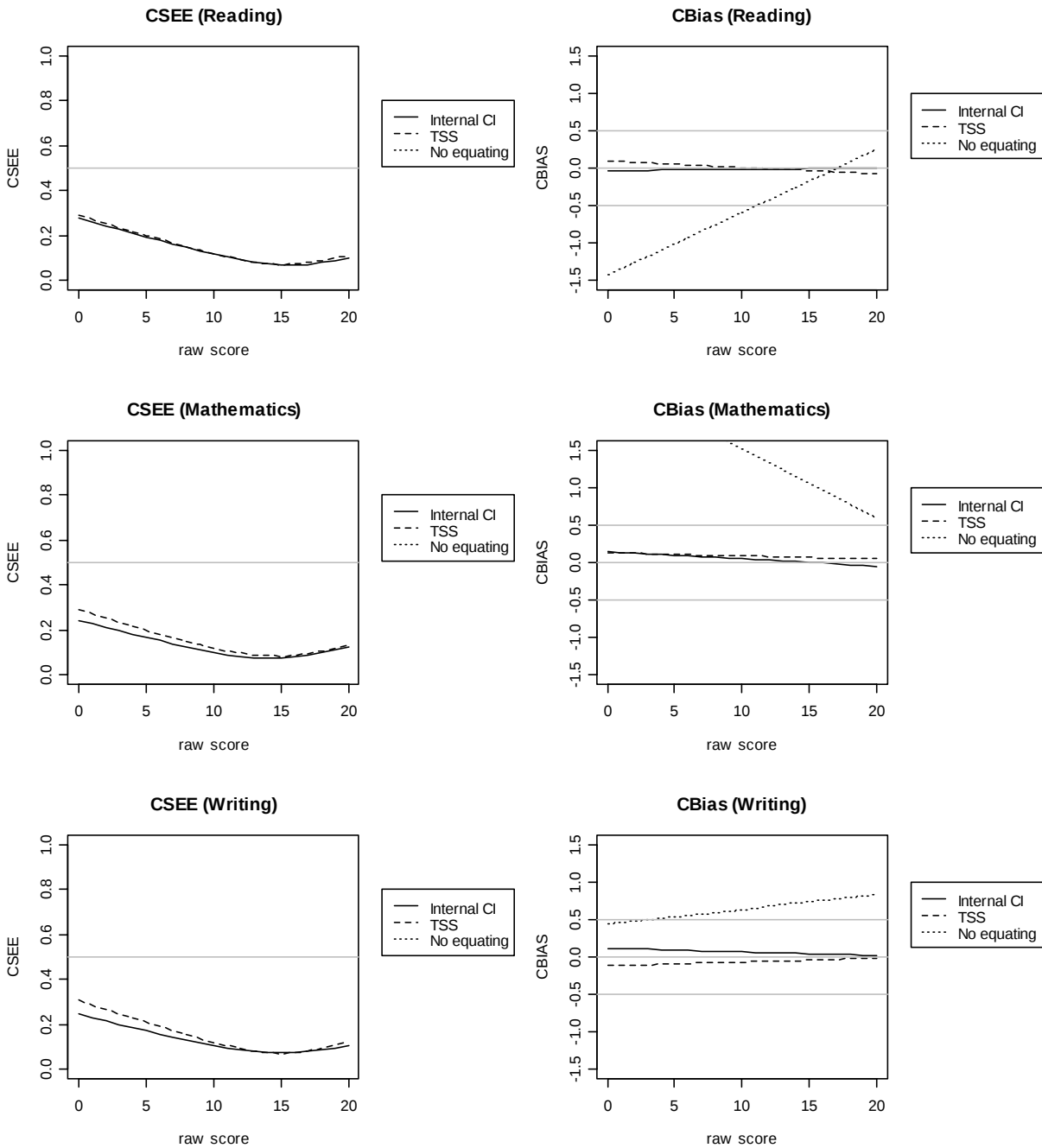


Figure 4. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, and writing subscore equatings for Test Y (Condition 1).

Note. The CSEE for no equating is zero and therefore not observed in the graph.

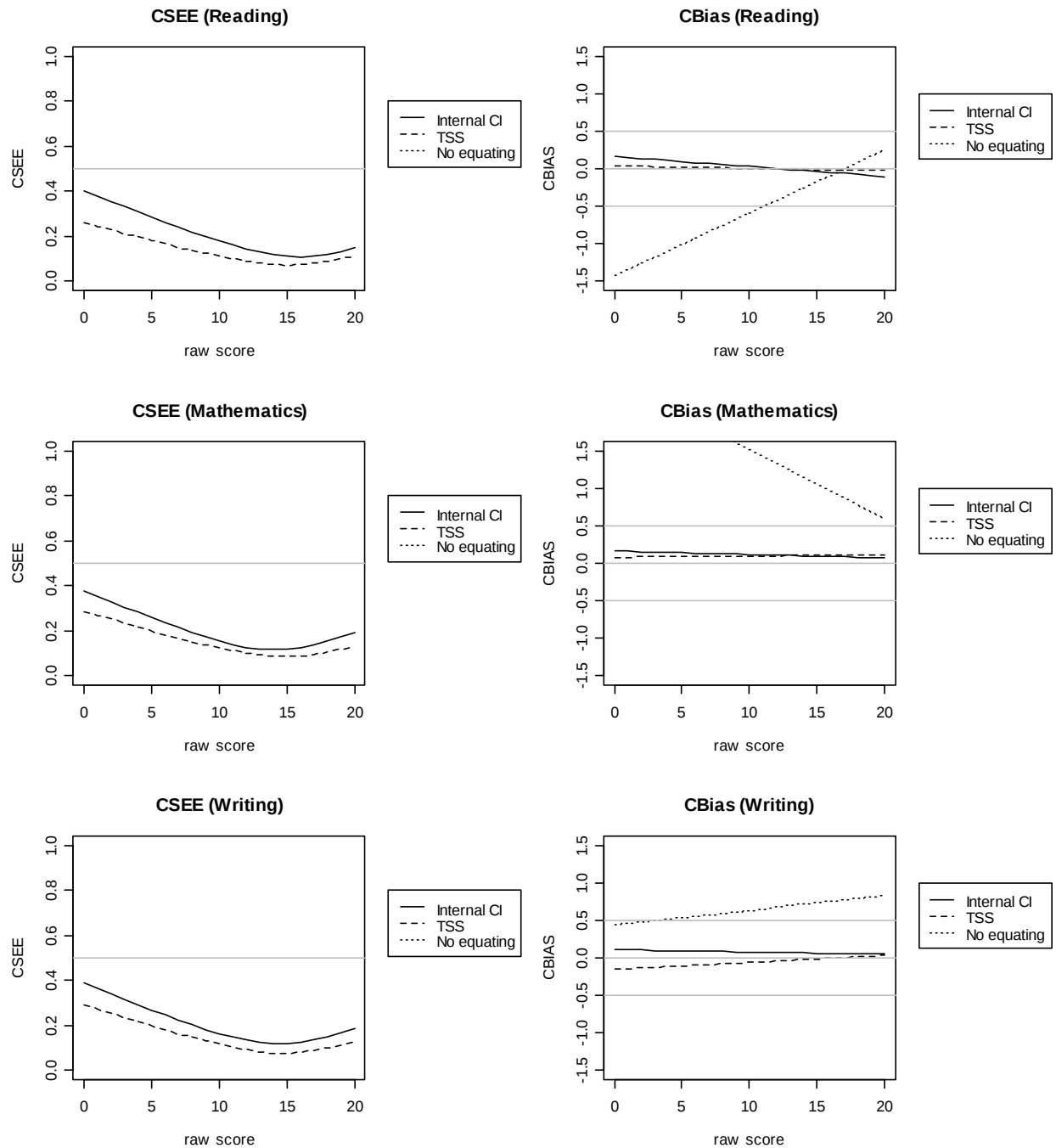


Figure 5. Conditional standard error of equating (CSEE) and conditional bias (CBias) for reading, math, and writing subscore equatings for Test Y (Condition 2).

Note. The CSEE for no equating is zero and therefore not observed in the graph.

quite large for the reading, social studies, and science subscores, suggesting that reporting subscores that are not equated may not be a reasonable choice for these subscores.

As seen in Figure 3 for Condition 2, the CSEEs for the internal common item equating and the TSS anchor equating for all the subscores were very similar to each other. The CSEEs (for most subscores) were also smaller than the DTM of 0.5 and could therefore be considered small. As seen in Figure 3, the CBias values for the internal common item equating and TSS anchor equating followed a similar pattern as was observed in Condition 1. They were quite small and fell within the DTM band for most of the subscores, except the reading subscore where the CBias values were slightly higher than the DTM for the internal common item equating in the lower score region. The Bias for no equating was quite large for the reading, social studies, and science subscores.

Results for Test Y

Average standard errors of equating (Avg SEE), weighted average bias (Bias), and root mean squared deviation (RMSD) results. As seen in Table 4, in Condition 1 (i.e., where a larger number of common items was used), the Avg SEE and RMSD for the subscore equatings (i.e., the reading, math, and writing subscores) that used internal common items as the anchor were always smaller than the subscore equatings that used TSS as an anchor. The Bias was also smaller for the reading and mathematics subscore equatings and about the same (in absolute magnitude) for the writing subscore equating. Similar to Test X, this was not completely unexpected because as seen in Table 2, the anchor-to-total correlations for each subscore were higher when common items as compared to TSS were used as the equating anchor.

However, in Condition 2, the pattern was reversed. The Avg SEEs for all the subscore equatings that used TSS as the anchor were smaller than the subscore equatings that used internal common items as the anchor. This finding was also true for the Bias and RMSD indexes. Considering the high anchor-to-total correlations for the TSS anchor equatings as compared to the internal common item equatings, this result seemed reasonable. Overall, the equatings using TSS as the anchor in this condition were more accurate (i.e., lower Avg SEE, Bias, and RMSD) for all the subscores.

Conditional standard errors of equating (CSEE) and weighted average bias (Bias). As seen in Figure 4 for Condition 1, the CSEEs for the internal common item equating were slightly smaller than the CSEEs for the TSS anchor equating for all the subscores at the lower score regions; they were very similar in the middle and higher score regions. Also, because the CSEEs for both subscore equatings were smaller than the DTM of 0.5, they could be considered

small. As seen in Figure 4, the CBias values for the internal common item equating and TSS anchor equating were quite small and within the DTM band for all the subscores. Similar to Test X, the Bias for no equating was quite large for all the subscores, suggesting that reporting subscores that are not equated may not be a reasonable choice for these subscores.

As seen in Figure 5 for Condition 2, the CSEEs for TSS anchor equatings were smaller than the internal common item equatings for all the subscores. The CSEEs for both the TSS and common item equatings were also smaller than the DTM of 0.5 and could therefore be considered small. Similar to Condition 1, the CBias values for the internal common item equating and TSS anchor equating were quite small and within the DTM band for all the subscores. However, the Bias for no equating was quite large for all the subscores.

Discussion and Conclusion

Although the demand for subscores is increasing, certain important factors must be considered before subscores can be considered useful. Comparability of subscores across different forms of a test is considered an important factor that must be achieved before subscores are reported for diagnostic, remedial, or other high stakes purposes. Although guidelines for equating total scores under different equating designs can be found in the equating literature, similar guidelines for equating subscores (especially in the classical equating context) are not well documented. Therefore the purpose of this study was to evaluate different approaches to effectively equate subscores from different forms of the same test. Subscore equatings conducted using either internal common items or TSS (that are comparable across different forms of a test) as the anchor were compared under the nonequivalent anchor test design (NEAT). The no-equating approach was also compared with the above two approaches for the purpose of showing that if different forms of a test are not parallel, then reporting raw unequated subscores can raise serious comparability issues across different forms of a test.

The results for Test X showed that when large numbers of common items (approximately 50% of the total score) were used to conduct the subscore equatings, then equating using internal common items as an anchor outperformed equating that used TSS as the anchor (i.e., lower Avg SEE, Bias, and RMSD). Similarly, the CSEE and CBias indexes were also lower for the subscore equatings that used internal common items as compared to TSS as an anchor, although in both situations the actual CSEE and CBias values were very small. These results could be interpreted by taking into consideration the anchor-to-total correlations in the

new and old forms. When the number of common items was large, the anchor-to-total correlation was higher if common items were used as the anchor as compared to TSS as the anchor, possibly resulting in a slightly more accurate equating. Finally, the no-equating approach did not seem reasonable for most subscores as it resulted in a considerable amount of equating bias and should not be considered as a viable option when the intent is to produce comparable subscores across different forms of a test. However, if the new and old forms are indeed parallel (which is virtually unknown in practice), then use of the unequated raw scores may result in little or no Bias, as was observed for the math subscore.

The results for Test X showed that when a smaller number of common items (approximately 25% of the total score which is common practice) was used to conduct the subscore equating, then equating using TSS as the anchor outperformed the equating using internal common items for three of the four subscores (i.e., lower RMSD). This finding may be attributed to the higher correlation observed between the subscore and the anchor when TSS was used. The CSEEs for both the internal common item and TSS anchor conditions were very similar, and the CBias values followed a similar pattern as the CBias values in Condition 1, except for the reading subscore where the CBias values were slightly larger in the lower score region for the equating that used internal common items as an anchor.

The results for Test Y showed that when a large number of common items (approximately 50% of the total score) was used to conduct the subscore equatings, then equating using internal common items as an anchor outperformed equating that used TSS as the anchor (i.e., lower Avg SEE, Bias, and RMSD). Similarly, the CSEE and CBias values were also lower for the subscore equatings that used internal common items as compared to TSS as the anchor, although in both cases the values were very small. Similar to Test X, when the number of common items was large, the anchor-to-total correlation was higher when common items were used as the anchor as compared to TSS as the anchor, resulting therefore in a slightly more accurate equating. The no-equating approach did not seem reasonable for the subscores, as it resulted in a considerable amount of equating bias.

The results for Test Y showed that when a smaller number of common items (approximately 25% of the total score, which is common practice) was used to conduct the subscore equating, the equating using TSS as the anchor outperformed equating using internal common items for all subscores (i.e., lower Avg SEE, Bias, and RMSD). The CSEEs for both the

internal common item and TSS anchor conditions were very similar, and the CBias values followed a pattern similar to the CBias values in Condition 1.

The results of this study suggest that when the number of common items in the new and old forms is large, the common item equating method may be the preferred method to equate subscores because they are expected to adequately represent the total subscore. However, if the correlation between the subscores and the TSS is at least moderately high (i.e., > 0.70), then using the TSS to equate the subscores may be a reasonable option. As seen in this study, although a large number of common items resulted in a more accurate equating in terms of lower equating error and bias, using a TSS as an anchor in those conditions also resulted in very small random equating error and bias (i.e., lower than the DTM for most subscores). Because it is not common practice to use a large number of common items in actual practice, the second condition where the number of common items is approximately 25% of the total may be more informative for testing programs interested in reporting equated subscores. The results from this condition suggest that when the number of common items is small, then using the TSS as the anchor is a good alternative for equating subscores because it results in a more precise equating as compared to common item equating using very few common items.

Future Research

Subscore equating was examined in this study for tests where the subscores were correlated at least moderately high with the total test (approximately 0.70 or higher). With such correlations, it is reasonable to expect the TSS to produce a fairly accurate equating. However, if the correlation between the total and subscores is low, then it is not expected that using the TSS as an anchor will produce an accurate equating. It should be noted, however, that under the same circumstances (i.e., low correlation between total and subscores), if very few common items are used as an anchor, then the common item approach may not produce an accurate equating either. Because we used real test data, it was not possible to manipulate the correlation between the subscore and the total score. Future studies may consider simulating test data where the correlation between the subscore and total score is low. If these studies find that using TSS or very few internal common items as an anchor produce inaccurate equatings, then in such cases the only viable option would be to increase the number of common items in the test to be able to effectively equate the subscores. However, this raises test security concerns. Therefore, testing programs must carefully evaluate the benefits of reporting equated subscores as compared to the

risk of overexposure of items before making a decision as to whether they want to incorporate more common items between old and new forms.

The subscores for Tests X and Y were relatively long (i.e., 21 items for each subscore on Test X and 20 items for each subscore on Test Y) in this study. Future studies should examine tests with fewer items (e.g., 10) in some or all of the subscores. Fewer items make it even more difficult to use the common item equating design for equating subscores. For example, consider a situation where the new and old forms have 25% overlap (common) at the total test level. If one of the subscores has 10 items, then the number of common items that may be available for the subscore equating could be fewer than three items, which may not adequately represent the content and difficulty of the subscore. In such situations, equating through the scaled total scores may be the only reasonable choice. Future subscore equating studies should also be conducted in a small sample context, where the subscore equatings can be affected not only by the small number of common items, but also by the small sample sizes.

We evaluated the feasibility in this study of using TSS as the anchor to equate subscores. Another possible option, which may be evaluated in a future study, is to use all the common items in the total test as the anchor to equate the subscores. This approach is expected to produce results similar to using TSS as the anchor but may have a practical advantage for testing programs with strict timelines.⁴ Unlike the TSS approach where the total scores have to be equated and scaled before they can be used as the anchor to conduct the subscore equatings, under the *all common item* approach, the subscore equatings can be conducted simultaneously or even before the total score is equated and scaled.

Only the chained linear and chained equipercentile methods were used in this study to conduct the subscore equatings. Future studies may use other methods such as Tucker, Levine, and kernel equating to evaluate whether the current results would generalize to other equating methods. Finally, the present study involved only two tests. Although it is reasonable to expect that the results of the current study will be similar if other similar tests were used (e.g., tests with similar correlations between the subscores and the total scores, similar content), more tests should be examined for greater generalizability of these results. In the meantime, results of this study may provide some guidelines for testing programs where a demand exists for subscore reporting and, therefore, the need to equate subscores.

References

- Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.
- Kim, S., Walker, M. E., & McHale, F. (2008). *Comparison among designs for equating constructed response tests* (ETS Research Rep. No. RR-08-53). Princeton, NJ: ETS.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.
- Haberman, S. J. (2008). *Subscores and validity* (ETS Research Rep. No. RR-08-64). Princeton, NJ: ETS.
- Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions*, 24(7), 349–368.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS
- Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement*, 30, 23–29.
- Puhan, G., Sinharay, S., Haberman, S. J., Larkin, K. (2010). Comparison of subscores based on classical test theory. *Applied Measurement in Education*, 23, 1-20
- Sinharay, S. (2010). When can subscores be expected to have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47(2), 150-174.
- Sinharay, S., Haberman, S., Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26(4), 21–28.
- Yao, L., & Boughton, K. A. (April, 2006). *Multidimensional linking of diagnostic subscale scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Wainer, H., Vevea, J.L., Camacho, F., Reeve, B, Rosa, K., Nelson, L., et al. (2001). Augmented scores—“Borrowing strength” to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Lawrence Erlbaum Associates.

Notes

- ¹ See Haberman, 2008; Sinharay et al., 2007; Wainer et al., 2001 for details on procedures proposed for increasing the reliability of subscores.
- ² Since equating subscores using total scaled scores (TSS) as an anchor somewhat deviates from the ideal characteristics of the anchor test (i.e., a mini test), this approach may be considered a linking rather than equating process. However, to simplify presentation, the term “equating” will be used throughout the paper. This approach may be considered similar to equating a constructed response (CR) test using a multiple-choice (MC) external anchor test, which although does not fulfill the ideal mini test requirement, may still produce accurate equatings as long as the CR and MC tests are moderately to highly correlated (see Kim, Walker, & McHale, 2008, for an example).
- ³ The chained equipercentile equating results can be obtained from the first author upon request.
- ⁴ We conducted some preliminary analyses that showed the equating bias was almost identical for both approaches, although random equating error was slightly smaller for the TSS approach.

Appendix

Indices to measure equating variability and bias

The formula for CSEE is

$$CSEE_j = \sqrt{\frac{1}{I} \sum_i \left(\hat{e}_y(x_{ij}) - \frac{\sum_i \hat{e}_y(x_{ij})}{I} \right)^2},$$

where I is the number of replications in the simulation ($i = 1$ to I , and $I = 500$) and $\hat{e}_y(x_{ij})$ is the equated score at score = x_j estimated for replication = i .

The formula for Avg SEE is

$$\text{Avg SEE} = \sqrt{\sum_j p_j CSEE_j^2},$$

where p_j is the raw proportion of examinees at each score point in the new form data.

The formula for CBias is

$$CBias_j = \frac{1}{I} \sum_i (\hat{e}_y(x_{ij}) - e_y(x_j)),$$

where I is the number of replications in the simulation ($i = 1$ to I , and $I = 500$), $e_y(x_j)$ is the criterion single group equated score at score = x_j and $\hat{e}_y(x_{ij})$ is the equated score at score = x_j estimated for replication = i .

The formula for Bias is

$$\text{Bias} = \sum_j p_j CBias_j,$$

where p_j is the raw proportion of examinees at each score point in the new form data.

The formula for RMSD is

$$RMSD = \sqrt{AvgBias^2 + AvgSEE^2},$$

where $AvgBias^2$ is the sum of the squared CBias values weighted by the raw proportion of examinees at each score point in the new form data. The formula is

$$AvgBias^2 = \sum_j p_j CBias_j^2.$$